

Regressão

Prof. André Gustavo Hochuli

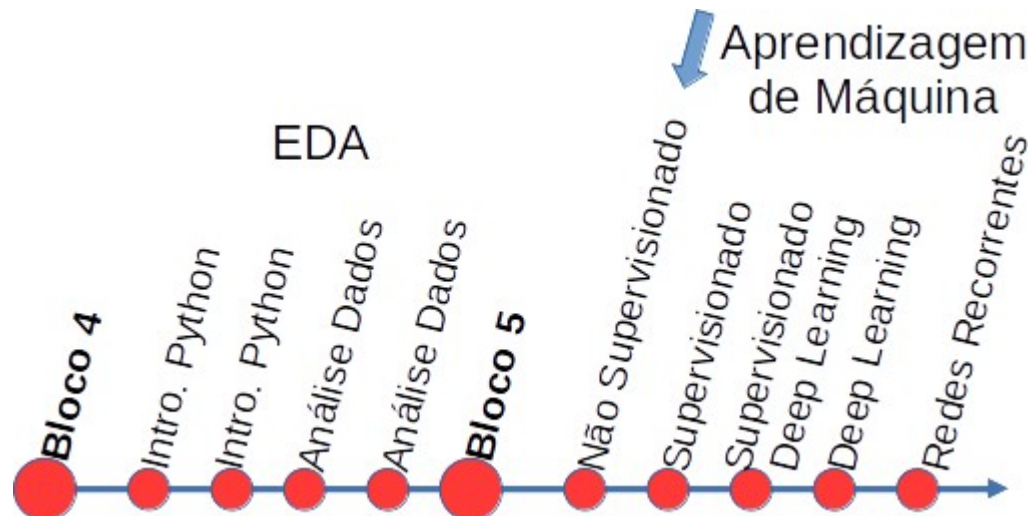
gustavo.hochuli@pucpr.br

aghochuli@ppgia.pucpr.br

github.com/andrehochuli/teaching

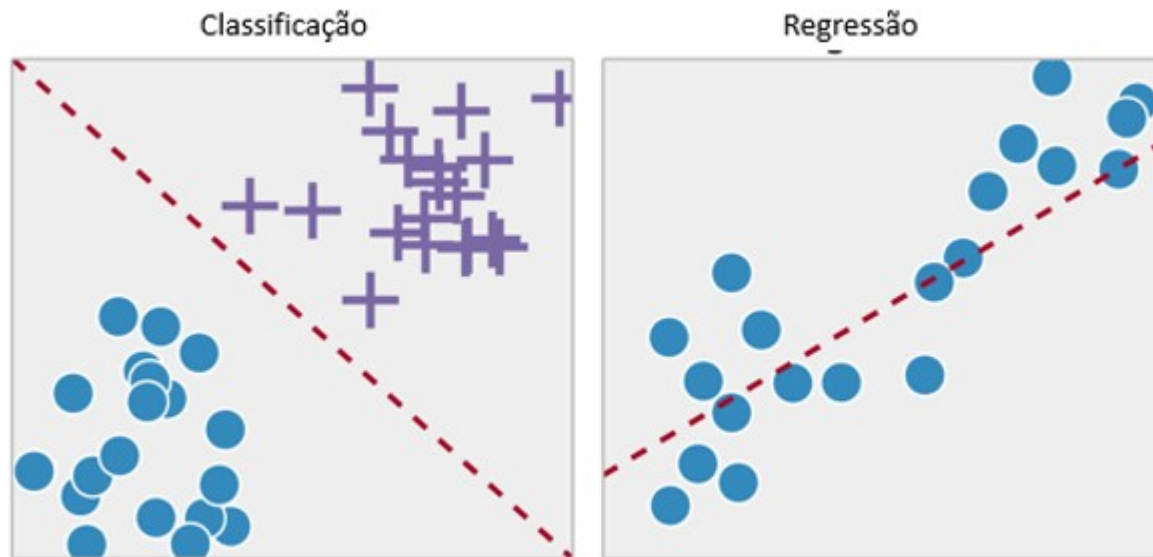
Plano de Aula

- Revisão Avaliação 1
- Regressão
- Exercícios



Regressão vs Classificação

- Classificação: Determina uma classe (0,1,2,3)
- Regressão: Determina valores contínuos (preço, clima, vendas, logística, sinais...)



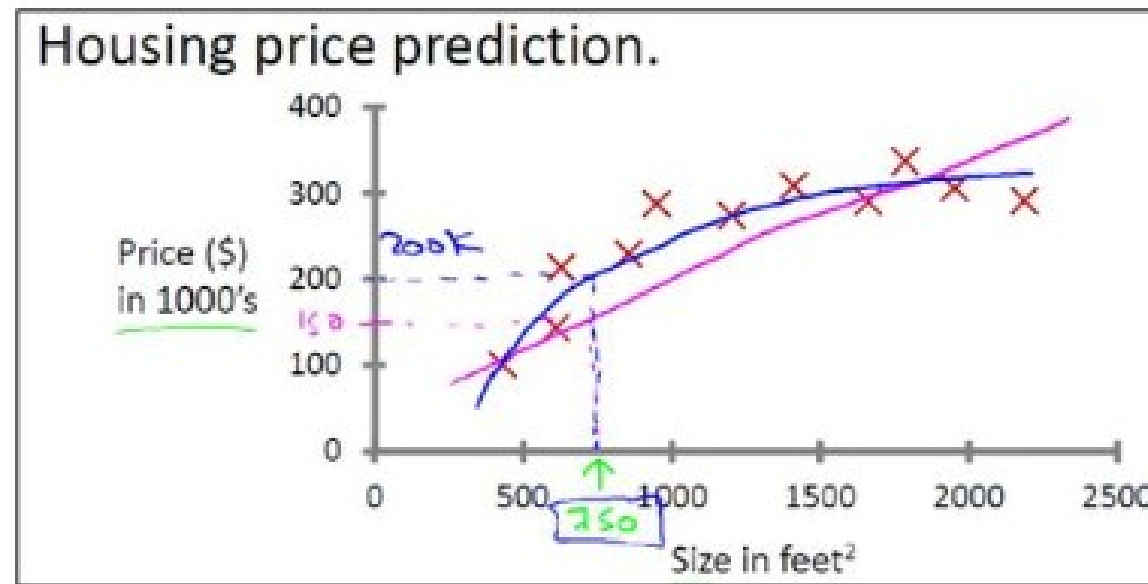
Regressão vs Classificação

Regressão: Compreender a relação entre as características dos dados e a variável dependente (target).

Mapeio X em Y contínuo

Linear

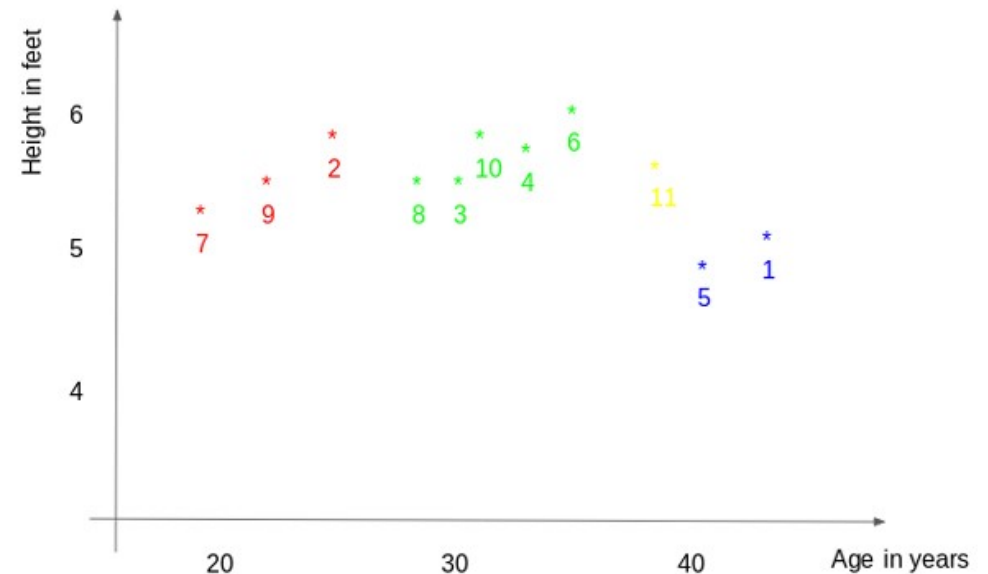
Não Linear



Regressão - KNN

Considere o dataset para determinar o peso de uma pessoa:

| ID | Height | Age | Weight |
|----|--------|-----|--------|
| 1 | 5 | 45 | 77 |
| 2 | 5.11 | 26 | 47 |
| 3 | 5.6 | 30 | 55 |
| 4 | 5.9 | 34 | 59 |
| 5 | 4.8 | 40 | 72 |
| 6 | 5.8 | 36 | 60 |
| 7 | 5.3 | 19 | 40 |
| 8 | 5.8 | 28 | 60 |
| 9 | 5.5 | 23 | 45 |
| 10 | 5.6 | 32 | 58 |
| 11 | 5.5 | 38 | ? |



Regressão - KNN

Dado K vizinhos, computa-se a média

K=3

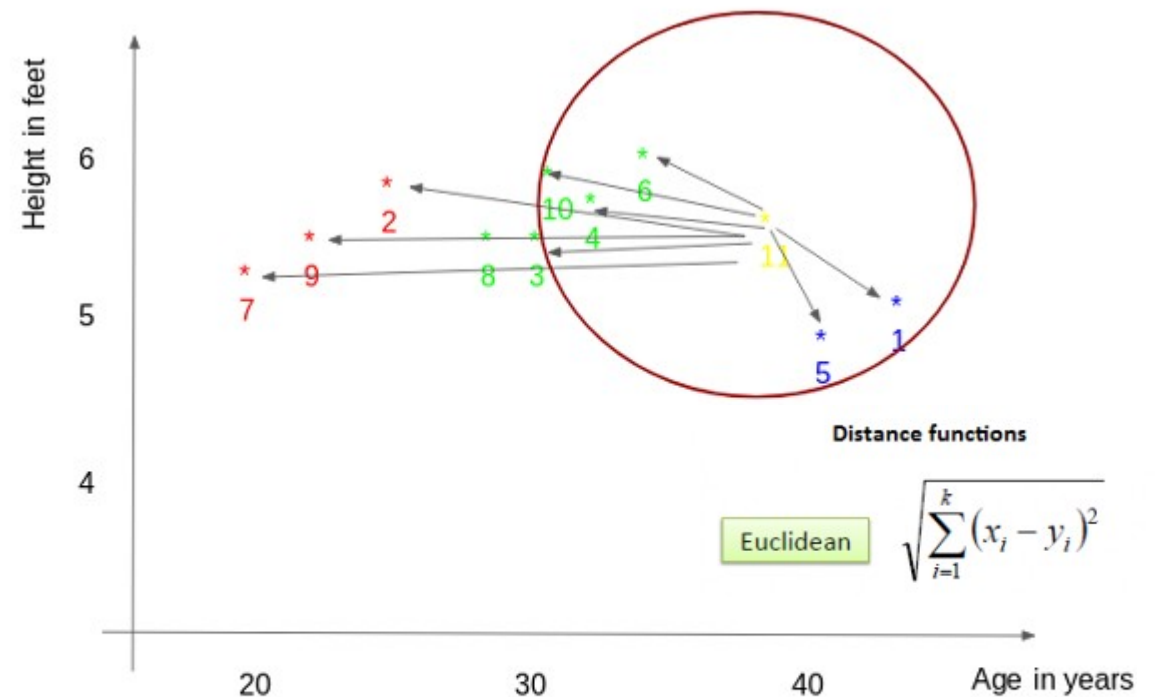
$$ID11 = (77+72+60)/3$$

$$ID11 = 69.66 \text{ kg}$$

K=50

$$ID\ 11 = (77+59+72+60+58)/5$$

$$ID\ 11 = 65.2 \text{ kg}$$



| ID | Height | Age | Weight |
|----|--------|-----|--------|
| 1 | 5 | 45 | 77 |
| 4 | 5.9 | 34 | 59 |
| 5 | 4.8 | 40 | 72 |
| 6 | 5.8 | 36 | 60 |
| 10 | 5.6 | 32 | 58 |

Regressão

Como avaliar o erro ?

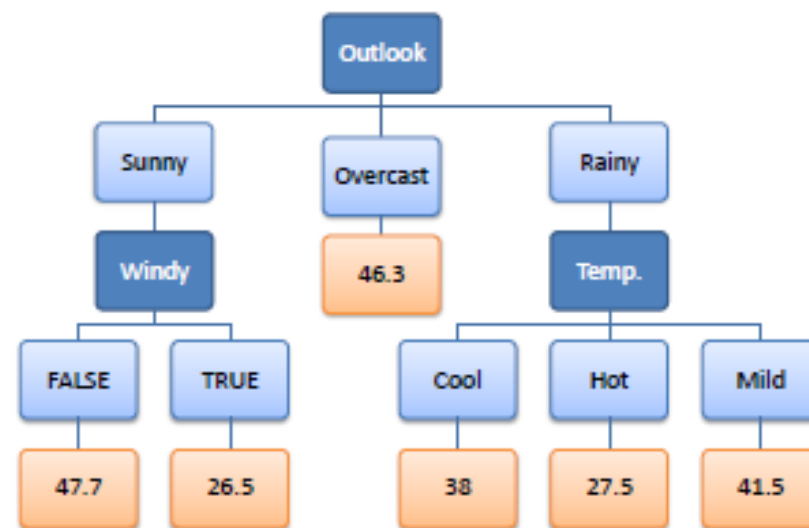
Mean-Square-Error (MSE)

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

| Height | Age | Weight | Predicted | Diff |
|--------|-----|--------|-----------|------|
| 5 | 40 | 71 | 69 | -2 |
| 5.9 | 23 | 86 | 80 | -6 |
| 4.5 | 51 | 65 | 77 | 12 |
| 5.3 | 36 | 89 | 82 | -7 |
| 5.1 | 48 | 75 | 67 | -8 |
| | | | MSE | 121 |

Regressão - Árvores

| Predictors | | | | Target |
|------------|-------|----------|-------|--------------|
| Outlook | Temp. | Humidity | Windy | Hours Played |
| Rainy | Hot | High | False | 26 |
| Rainy | Hot | High | True | 30 |
| Overcast | Hot | High | False | 48 |
| Sunny | Mild | High | False | 46 |
| Sunny | Cool | Normal | False | 62 |
| Sunny | Cool | Normal | True | 23 |
| Overcast | Cool | Normal | True | 43 |
| Rainy | Mild | High | False | 36 |
| Rainy | Cool | Normal | False | 38 |
| Sunny | Mild | Normal | False | 48 |
| Rainy | Mild | Normal | True | 48 |
| Overcast | Mild | High | True | 62 |
| Overcast | Hot | Normal | False | 44 |
| Sunny | Mild | High | True | 30 |



Regressão - Árvores

Determina a homogeneidade pelo do desvio padrão, média e coeficiente de variação

| Hours Played |
|--------------|
| 25 |
| 30 |
| 46 |
| 45 |
| 52 |
| 23 |
| 43 |
| 35 |
| 38 |
| 46 |
| 48 |
| 52 |
| 44 |
| 30 |



$$\text{Count} = n = 14$$

$$\text{Average} = \bar{x} = \frac{\sum x}{n} = 39.8$$

$$\text{Standard Deviation} = S = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = 9.32$$

$$\text{Coefficient of Variation} = CV = \frac{S}{\bar{x}} * 100\% = 23\%$$

Regressão - Árvores

Desvio padrão para dois atributos:

$$S(T, X) = \sum_{c \in X} P(c)S(c)$$

| | | Hours Played (StDev) | Count |
|---------|----------|----------------------------|-------|
| Outlook | Overcast | 3.49 | 4 |
| | Rainy | 7.78 | 5 |
| | Sunny | 10.87 | 5 |
| | | | 14 |



$$\begin{aligned} S(\text{Hours}, \text{Outlook}) &= P(\text{Sunny}) * S(\text{Sunny}) + P(\text{Overcast}) * S(\text{Overcast}) + P(\text{Rainy}) * S(\text{Rainy}) \\ &= (4/14) * 3.49 + (5/14) * 7.78 + (5/14) * 10.87 \\ &= 7.66 \end{aligned}$$

Regressão

$$SDR(T, X) = S(T) - S(T, X)$$

Redução do Desvio Padrão

$$\begin{aligned} \text{SDR}(\text{Hours}, \text{Outlook}) &= S(\text{Hours}) - S(\text{Hours}, \text{Outlook}) \\ &= 9.32 - 7.66 = 1.66 \end{aligned}$$

O maior SDR é escolhido como raiz

| | | Hours Played (StDev) |
|---------|----------|----------------------|
| Outlook | Overcast | 3.49 |
| | Rainy | 7.78 |
| | Sunny | 10.87 |
| | | SDR=1.66 |


| | | Hours Played (StDev) |
|-------|------|----------------------|
| Temp. | Cool | 10.51 |
| | Hot | 8.95 |
| | Mild | 7.65 |
| | | SDR= 0.48 |

| | | Hours Played (StDev) |
|----------|--------|----------------------|
| Humidity | High | 9.36 |
| | Normal | 8.37 |
| | | SDR=0.28 |

| | | Hours Played (StDev) |
|-------|-------|----------------------|
| Windy | False | 7.87 |
| | True | 10.59 |
| | | SDR=0.29 |

Regressão

Como escolher o critério de parada?



A decision tree diagram with a root node 'Outlook' branching into three categories: 'Sunny', 'Overcast', and 'Rainy'. Each category is associated with a table of data points.

| Outlook | Temp | Humidity | Windy | Hours Played |
|---------|------|----------|-------|--------------|
| Sunny | Mild | High | FALSE | 45 |
| Sunny | Cool | Normal | FALSE | 52 |
| Sunny | Cool | Normal | TRUE | 23 |
| Sunny | Mild | Normal | FALSE | 46 |
| Sunny | Mild | High | TRUE | 30 |

| Overcast | Temp | Humidity | Windy | Hours Played |
|----------|------|----------|-------|--------------|
| Overcast | Hot | High | FALSE | 46 |
| Overcast | Cool | Normal | TRUE | 43 |
| Overcast | Mild | High | TRUE | 52 |
| Overcast | Hot | Normal | FALSE | 44 |

| Rainy | Temp | Humidity | Windy | Hours Played |
|-------|------|----------|-------|--------------|
| Rainy | Hot | High | FALSE | 25 |
| Rainy | Hot | High | TRUE | 30 |
| Rainy | Mild | High | FALSE | 35 |
| Rainy | Cool | Normal | FALSE | 38 |
| Rainy | Mild | Normal | TRUE | 48 |

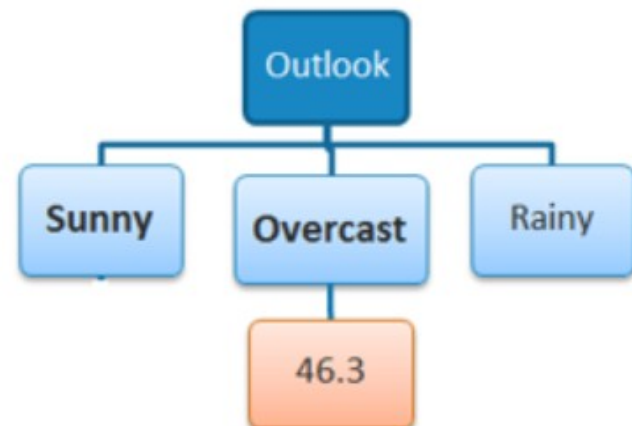
Regressão

Considere um limiar de $CV < 10\%$ ou $\text{count} \leq 3$

Overcast => OK

Rainy e Sunny ainda precisam de 'divisões'

| | | Hours Played (StDev) | Hours Played (AVG) | Hours Played (CV) | Count |
|---------|----------|----------------------------|--------------------------|-------------------------|-------|
| Outlook | Overcast | 3.49 | 46.3 | 8% | 4 |
| | Rainy | 7.78 | 35.2 | 22% | 5 |
| | Sunny | 10.87 | 39.2 | 28% | 5 |



Regressão

Sunny:

‘Windy’ é determinante

Maior SDR

CV > 8% ou count <= 3

Outlook - Sunny

| Temp | Humidity | Windy | Hours Played |
|------|----------|-------|--------------|
| Mild | High | FALSE | 45 |
| Cool | Normal | FALSE | 52 |
| Cool | Normal | TRUE | 23 |
| Mild | Normal | FALSE | 46 |
| Mild | High | TRUE | 30 |
| | | | S = 10.87 |
| | | | AVG = 39.2 |
| | | | CV = 28% |

| | | Hours Played (StDev) | Count |
|------|------|----------------------|-------|
| Temp | Cool | 14.50 | 2 |
| | Mild | 7.32 | 3 |

$$SDR = 10.87 - ((2/5) * 14.5 + (3/5) * 7.32) = 0.678$$

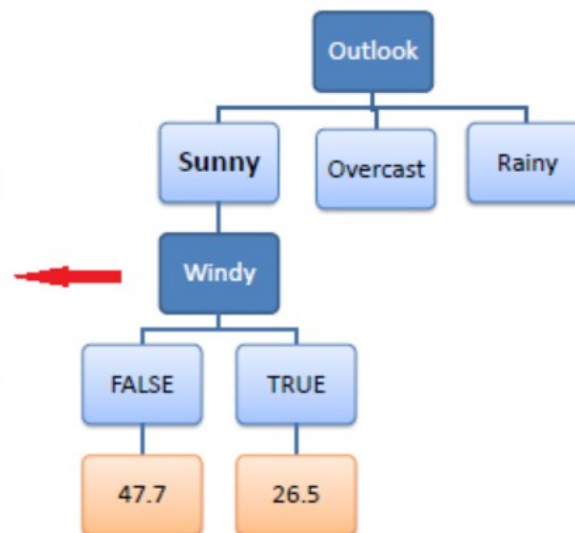
| | | Hours Played (StDev) | Count |
|----------|--------|----------------------|-------|
| Humidity | High | 7.50 | 2 |
| | Normal | 12.50 | 3 |

$$SDR = 10.87 - ((2/5) * 7.5 + (3/5) * 12.5) = 0.370$$

| | | Hours Played (StDev) | Count |
|-------|-------|----------------------|-------|
| Windy | False | 3.09 | 3 |
| | True | 3.50 | 2 |

$$SDR = 10.87 - ((3/5) * 3.09 + (2/5) * 3.5) = 7.62$$

| Temp | Humidity | Windy | Hours Played |
|------|----------|-------|--------------|
| Mild | High | FALSE | 45 |
| Cool | Normal | FALSE | 52 |
| Mild | Normal | FALSE | 46 |
| Cool | Normal | TRUE | 23 |
| Mild | High | TRUE | 30 |



Regressão

Rainy:

'Temp' é determinante

Maior SDR

CV > 8% ou count <= 3

Outlook - Rainy

| Temp | Humidity | Windy | Hours Played |
|------|----------|-------|--------------|
| Hot | High | FALSE | 25 |
| Hot | High | TRUE | 30 |
| Mild | High | FALSE | 35 |
| Cool | Normal | FALSE | 38 |
| Mild | Normal | TRUE | 48 |
| | | | $S = 7.78$ |
| | | | AVG = 35.2 |
| | | | CV = 22% |

| | | Hours Played (StDev) | Count |
|------|------|----------------------|-------|
| Temp | Cool | 0 | 1 |
| | Hot | 2.5 | 2 |
| | Mild | 6.5 | 2 |

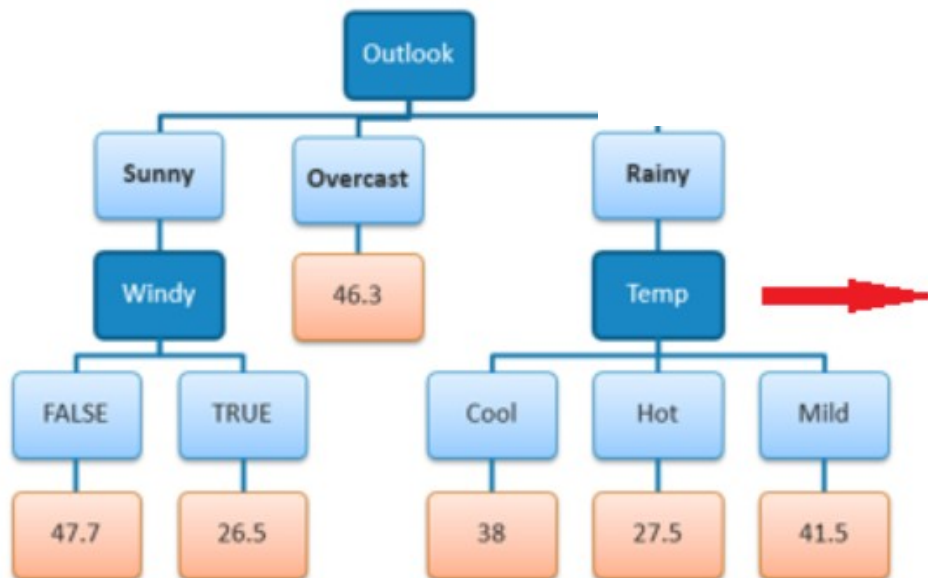
$$SDR = 7.78 - ((1/5)*0 + (2/5)*2.5 + (2/5)*6.5) = 4.18$$

| | | Hours Played (StDev) | Count |
|----------|--------|----------------------|-------|
| Humidity | High | 4.1 | 3 |
| | Normal | 5.0 | 2 |

$$SDR = 7.78 - ((3/5)*4.1 + (2/5)*5.0) = 3.32$$

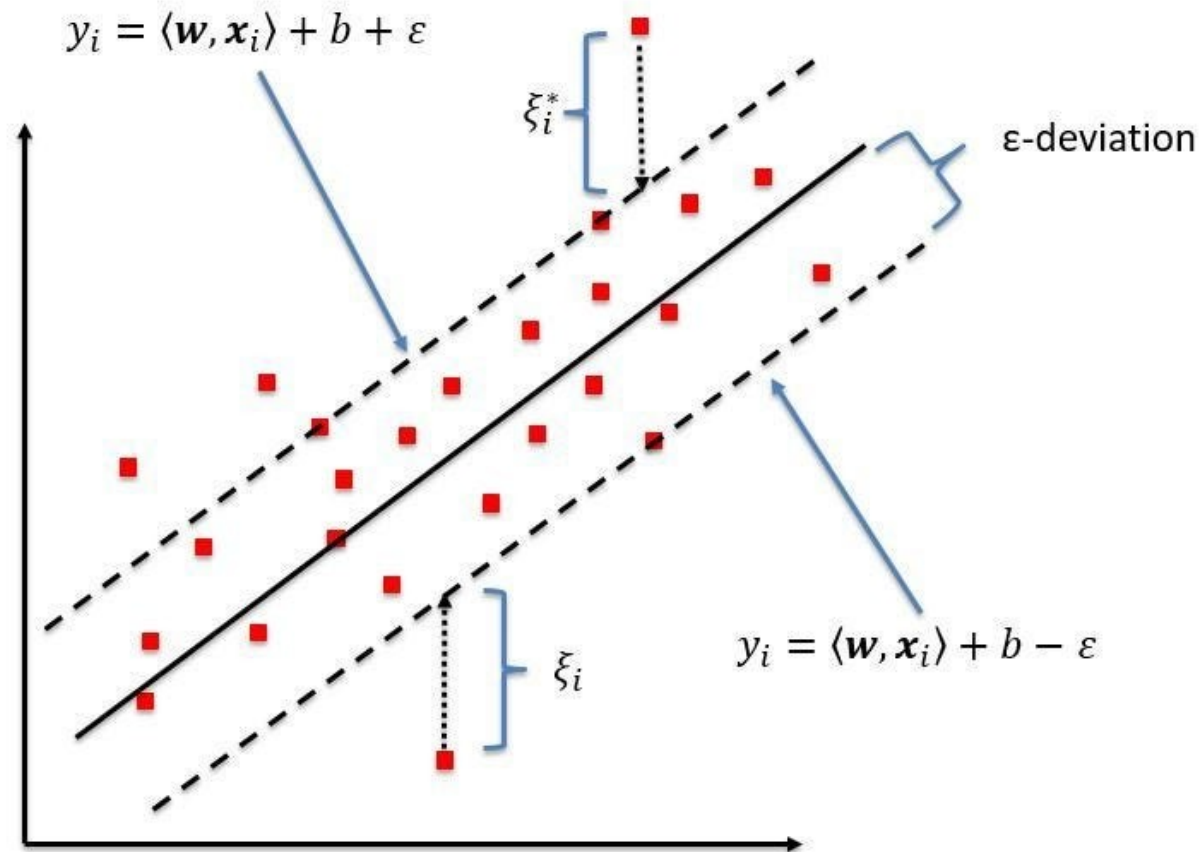
| | | Hours Played (StDev) | Count |
|-------|-------|----------------------|-------|
| Windy | False | 5.6 | 3 |
| | True | 9.0 | 2 |

$$SDR = 7.78 - ((3/5)*5.6 + (2/5)*9.0) = 0.82$$



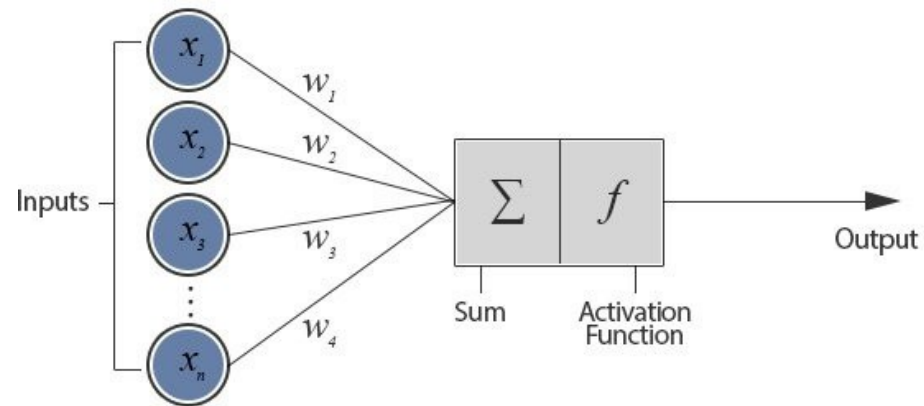
| Temp | Hours Played |
|------|--------------|
| Cool | 38 |
| Hot | 25 |
| Hot | 30 |
| Mild | 35 |
| Mild | 48 |

Regressão - SVM

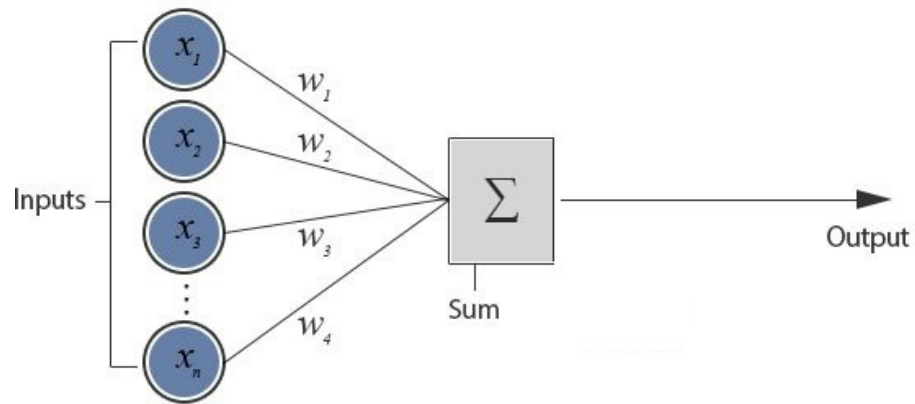


Regressão - MLP

Perceptron - Classificação

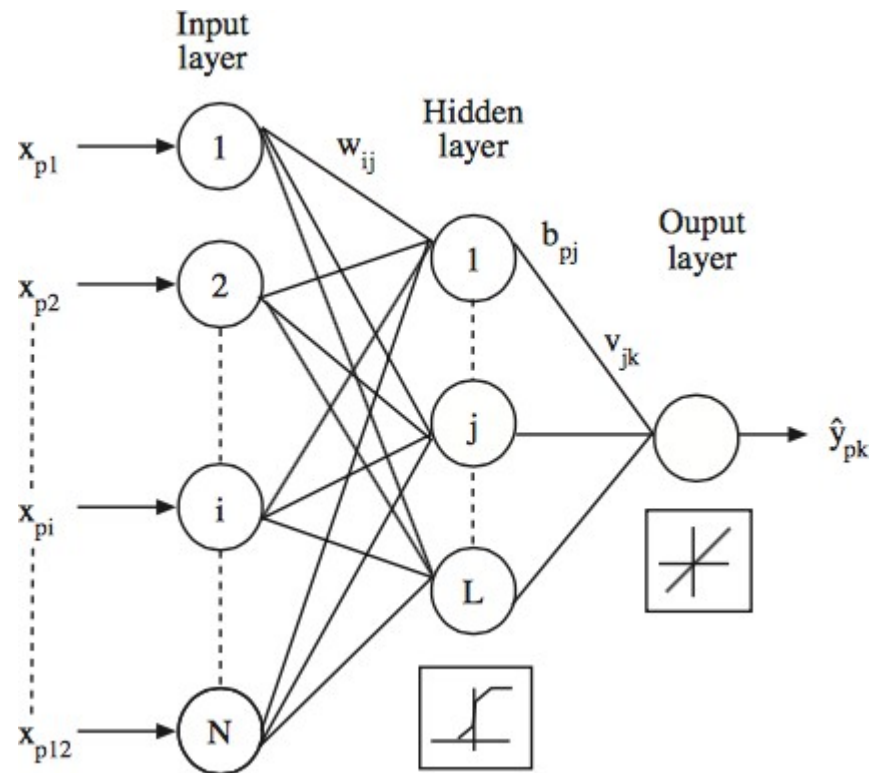


Perceptron - Regressão



Regressão - MLP

MLP - Classificação



Lets Code

Abordaremos a construção dos modelos de regressão utilizando um dataset para prever preços de casas.

Acompanhe e crie sua implementação em conjunto com o professor.

Código base em: [Tópico 03 - Regressão.ipynb](#)